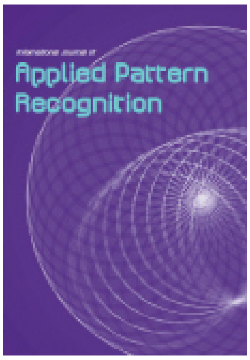


International Journal of Applied Pattern Recognition

ISSN online: 2049-8888 - ISSN print: 2049-887X
https://www.inderscience.com/ijapr

COVID-19's X-ray images classification: training from scratch or transfer learning?

Eliú Moreno-Ramírez, Héctor Anaya-Sánchez, José Fco. Martínez-Trinidad, J. Ariel Carrasco-Ochoa

DOI: 10.1504/IJAPR.2024.10071294

Article History:

Received:	13 September 2024
Last revised:	06 March 2025
Accepted:	10 April 2025
Published online:	19 June 2025

COVID-19’s X-ray images classification: training from scratch or transfer learning?

Eliú Moreno-Ramírez,
Héctor Anaya-Sánchez*,
José Fco. Martínez-Trinidad and
J. Ariel Carrasco-Ochoa

Ciencias Computacionales,
Instituto Nacional de Astrofísica Óptica y Electrónica,
Luis Enrique Erro #1, Sta María Tonanzintla,
Cholula, 72840, Puebla, Mexico
Email: eliu.moreno@inaoep.mx
Email: hector.anaya@inaoep.mx
Email: fmartine@inaoep.mx
Email: ariel@inaoep.mx
*Corresponding author

Abstract: This paper presents a comprehensive analysis of 11 state-of-the-art deep convolutional neural network (CNN) models for COVID-19’s X-ray image classification with the two configurations more studied in the literature: transfer learning with fine-tuning and training from scratch. All models were assessed under the same experimental framework. Unlike other works, we used a dataset compiled from several public datasets, increasing its variability to reduce the risk of overfitting. Our results show which deep convolutional neural networks performed the best in accuracy and F1-score when training from scratch and with transfer learning.

Keywords: COVID-19 classification; deep learning; machine learning; transfer learning.

Reference to this paper should be made as follows: Moreno-Ramírez, E., Anaya-Sánchez, H., Martínez-Trinidad, J.F. and Carrasco-Ochoa, J.A. (2024) ‘COVID-19’s X-ray images classification: training from scratch or transfer learning?’, *Int. J. Applied Pattern Recognition*, Vol. 7, Nos. 3/4, pp.205–221.

Biographical notes: Eliú Moreno-Ramírez received his Master’s in Computer Science in 2024 at the Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), where he conducted research in the field of artificial intelligence. His main academic and scientific interests include machine learning, pattern recognition, cryptography, and reconfigurable computing.

Héctor Anaya-Sánchez received his Master’s in Computer Science from the Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) in 2024. He is currently a PhD student, where his main research topics are machine learning, pattern recognition and image generation.

José Fco. Martínez-Trinidad received his PhD in Computer Science from the Computing Research Center at the National Polytechnic Institute of Mexico in 2000. He is a researcher in the Computer Science Department at the Instituto Nacional de Astrofísica, Óptica y Electrónica of Mexico, and First Vice-President of the International Association for Pattern Recognition during 2024-2026. He has been a guest editor for several JCR Journals and a Pattern Recognition Journal associate editor. His work on pattern recognition for mixed data on fundamental problems and applications has been published in over 250 journal papers.

J. Ariel Carrasco-Ochoa received his PhD in Computer Science from the Computing Research Center of the National Polytechnic Institute (CIC-IPN) of Mexico, in 2001. He is currently the Head of the Department of Computational Sciences at the Instituto Nacional de Astrofísica, Óptica y Electrónica of Mexico. He has published more than 200 papers on topics related to pattern recognition, data mining, testor theory, feature and prototype selection, document analysis, and clustering.

1 Introduction

The ongoing COVID-19 pandemic, produced by the SARS-CoV-2 virus, is highly transmissible from person to person, even before clinical signs begin to show Wang et al. (2022). The most common clinical signs of COVID-19, as reported by Eskandarian et al. (2023), are dry cough, fever, diarrhea, sore throat, dyspnea, myalgia, shortness of breath, and bilateral lung infiltrates that can be observed on chest X-ray (CXR). Patients with severe COVID-19 have developed critical complications, such as septic shock, pulmonary edema, cardiac injury, acute respiratory distress syndrome, and even death (Kassania et al., 2021). These factors emphasise the importance of early detection of COVID-19 to provide proper patient care and control the spread of infection.

Despite the availability of vaccines, the world is still grappling with the consequences of recurrent infectious waves (Paul et al., 2023). Therefore, it is important to curb the spread of COVID-19 and prevent transmission by detecting patients early. X-ray imaging is an additional screening tool to help improve detection rates. Chest X-ray reveals radiological patterns specific to COVID-19, distinguished into ground-glass and mixed attenuation (Wu et al., 2021). Raw images often differ from COVID-19 radiographic patterns, as they typically contain prominent, well-defined objects, in contrast to the indistinct lung markings, opacity, and consolidation commonly observed in COVID-19 cases.

Deep convolutional neural network (CNN) architectures are powerful tools for analysing medical images, including COVID-19 detection (Aslani and Jacob, 2023; Bhosale and Patnaik, 2023). Several works have been developed to discriminate and identify COVID-19 through the analysis of X-ray images. In the literature, we can find works where deep CNN have been applied to COVID-19's X-ray images classification where the networks are trained from scratch. For instance, in Ramadhan and Baykara (2022) used a VGG16 CNN; they reduced the number of parameters in the architecture to work on cropped images (including only the chest area) for multiple classifications (three classes: COVID-19, normal, and pneumonia) and binary classification (COVID-19 and normal). By splitting the datasets into training and testing sets (70% and 30%, respectively), the proposed VGG16 trained from scratch obtained 97.50% accuracy for multiple classifications and 99.76% for binary classification. However, the highest number of images of COVID-19 used as training was 1,140. While Wang et al. (2022) proposed to use deep-learning-based approaches to classify COVID-19 and normal (healthy) CXR images. The authors designed a COVID-19 X-ray image detection model by applying a multi-head self-attention mechanism to the ResNet50V2 network bottleneck layer. The experimental results on a dataset with three categories (COVID-19, common pneumonia, and normal lungs) show that by training from scratch, the multi-head self-attention residual network (MHSA-ResNet) detection model has an accuracy of 95.52% and a precision of 96.02%. The number of images of COVID-19 used as training was 961. More recent works following this approach are Ukwuoma et al. (2023) and Abdullah et al. (2024). Unlike previous works that rely on a single dataset source with limited images of CXRs for the training phase and evaluation, these works used a more extensive and diverse dataset collected from different repositories and studies.

Transfer learning is a deep learning approach that involves training an architecture for a specific task and then using the trained architecture for another task. The approach relies on a pre-trained deep learning architecture created and fine-tuned by deep learning researchers using thousands or millions of sample images (Alsattar et al., 2024; Hassan et al., 2024), typically from the ImageNet domain, and fine-tuned for a new dataset. This approach is beneficial when data is insufficient to create a trained architecture from scratch. For instance, the authors from Farooq and Hafeez (2020) presented COVID-ResNet for classifying COVID-19 using fine-tuning, using ImageNet weights with a dataset of 2839 X-ray images. This CNN architecture achieved an accuracy of 96.23%. However, the authors commented that to make COVID-ResNet clinically applicable, it is crucial to train the architecture with a more comprehensive dataset and evaluate it on a larger cohort in real-world scenarios.

Kassania et al. (2021) conducted a study on applying machine learning (ML) algorithms to diagnose COVID-19 from X-ray and CT images. The proposed methodology first pre-processed a public dataset, consisting of X-ray and CT images of COVID-19, pneumonia, and healthy cases taken from Mooney (2018) and Stein et al. (2018). The pre-processing involved standard image normalisation techniques, followed by feature extraction using state-of-the-art deep CNN architectures to extract features of each input image. Different robust deep CNN architectures, such as MobileNet, DenseNet, Xception, InceptionV3, InceptionResNet50V2, ResNet50V2, VGGNet, and NASNet, were selected for their feasibility of transferring learning into computer vision. The output of these architectures (feature vectors) was then introduced into ML classifiers such as random forest, XGBoost, AdaBoost, Bagging, LightGBM, and decision tree. The study showed that the best accuracy was 99.00% ± 0.09, achieved by the Bagging classifier on features extracted by the DenseNet121 architecture.

Khan et al. (2022) used a dataset with 5,223 X-ray images of COVID-19 patients and, 9,904 non-COVID-19 patients. They employed different subsets, including 3,000 and 5,000 images of both classes and the entire dataset. The authors divided the dataset into 80% for training and 20% for testing; thus, the maximum number of images of COVID-19 used as training was around 4,178. For each phase, they used preprocessing techniques such as normalisation and data augmentation (rotations, translations, and zoom) and proposed two deep CNN architectures for discriminating between COVID-19 and healthy individuals. These architectures focused on extracting radiological patterns from X-rays using a convolution block based on split-transform merge. Then, they used channel boosting Khan et al. (2022) with transfer learning for better performance. With their proposal, they achieved an accuracy of 96.53%. This result was compared with other architectures such as InceptionV3, AlexNet, DenseNet201, Xception, VGG-16, and ResNet50V2.

Reddy et al. (2023) proposed a multimodal fusion transfer learning (MMF-DTL) model to classify COVID-19 from CXRs. It uses three pre-trained models (VGG16, InceptionV3, and ResNet50) for feature extraction, improving the detection rate by fusing these models. The softmax classifier categorises the images into six different classes. Experimental results on the CXR dataset show the model's high effectiveness, with a sensitivity of 92.96%, specificity of 98.54%, and precision of 93.60%.

Khattab et al. (2024) evaluated four pre-trained CNN models under the transfer learning approach to classify pneumonia in CXR images using four different datasets. Cross-entropy (CE) loss functions were employed for balanced datasets, and both CE and focal loss (FL) functions were used for imbalanced datasets. The results showed that InceptionResNet V2 achieved an accuracy of 88.63% in multi-class classification for the first dataset, while InceptionV3 achieved 94.35% and 97.67% in the second and fourth datasets, respectively. Xception stood out with a 100.00% accuracy in binary classification for the third dataset.

The works mentioned above allow us to appreciate that when transfer learning is used for COVID-19's X-ray image classification, the deep CNN architectures are pre-trained using ImageNet weights (Deng et al., 2009). According to Khan et al. (2022), there are many studies for the classification of patients with COVID-19, however, most of the studies have been developed with the use of CNN architectures which were trained with natural images. Additionally, fine-tuning is performed with a small dataset, as in the case of several works training the network from scratch.

Recent studies have continued to enhance our understanding of deep learning applications for COVID-19 detection. For instance, a 2023 study evaluated five deep learning models – ResNet50, ResNet101, DenseNet121, DenseNet169, and InceptionV3 – using transfer learning to identify COVID-19 from CXR images, with ResNet101 achieving the highest accuracy of 96% (Constantinou et al., 2023). Another 2024 study introduced a deep learning framework for accurate COVID-19 classification and severity prediction using CXR images, emphasising its potential in enhancing diagnostic precision and patient management (Singh et al., 2024). Furthermore, a retrospective study examined the generalisation capabilities of deep learning algorithms across different datasets, underscoring the importance of model robustness for clinical applications (Fernández-Miranda et al., 2024). Additionally, researchers assessed the performance of deep learning models such as VGG16, VGG19, DenseNet121, and ResNet50, demonstrating effective classification of COVID-19 and non-COVID-19 pneumonia cases (Cao et al., 2023).

As seen from the above paragraphs, researchers have applied various deep CNNs to address the problem of COVID-19's X-ray image classification. However, to our knowledge, a comparison of the different deep CNN architectures using transfer learning or training from scratch, under the same experimental framework has not been studied. In addition, several of the reported works commented that the studies were carried out with a small image dataset from a single source and, therefore, with little variability of the same. Thus, the motivation of this paper is to perform a comprehensive analysis of 11 state-of-the-art deep CNN models for COVID-19's X-ray image classification with two configurations: transfer learning with fine-tuning and scratch learning under the same experimental framework by employing the COVID-QU-Ex dataset, one of the most extensive and up-to-date datasets publicly available which were collected from different repositories and studies (Tahir et al., 2022). The deep CNN architectures studied are the most widely used in the literature: ResNet50V2 (He et al., 2016), InceptionV3 (Dongmei et al., 2020), Xception (Fabien, 2019), MobileNetV2 (Sandler et al., 2018), MobileNetV3 (Howard et al., 2019), VGG16, VGG19 (Simonyan and Zisserman, 2014), DenseNet121 (Huang et al., 2017), NASNet (Zoph et al., 2018) and EfficientB0 (Tan and Le, 2019).

This paper is divided into five sections. Section 2 describes the deep CNNs employed in our study for COVID-19's X-ray image classification. Section 3 presents the materials and methods used in our study. Section 4 provides the results and discusses them. Finally, Section 5 shows conclusions and some directions for future work.

2 Deep CNN architectures

As commented in the previous section, the problem of COVID-19's X-ray classification has been widely addressed, and several deep CNN architectures have been developed to solve it. Some studies have employed training from scratch with limited data, while others have utilised transfer learning by fine-tuning pre-trained architectures with limited data. Therefore, the deep CNN architectures used in our study are briefly described in the following section. Residual neural network (ResNet50V2) (He et al., 2016) is a deep CNN architecture that introduced the concept of residual learning. In ResNet50V2, residual blocks are used to overcome the problem of vanishing gradients, which occurs when training deep neural networks. A residual block is essentially a shortcut connection that bypasses one or more layers, allowing the gradient to flow directly from the input to the output of the block. This approach helps to maintain the gradient magnitude, making it easier to train deep networks. The ResNet50V2 architecture has achieved state-of-the-art performance on various image classification tasks, including the ImageNet large scale visual recognition challenge (ILSVRC).

MobileNet (Howard et al., 2017) is a CNN architecture designed to be more computationally efficient for mobile and embedded devices while maintaining good accuracy on image classification tasks. It achieves this efficiency by using depthwise separable convolutions, which factorise a standard convolution into a depthwise convolution and a pointwise convolution, reducing the computational cost. MobileNet also uses linear bottleneck layers and global average pooling to reduce further the number of parameters and computations needed. The first version, MobileNetV1, was introduced in 2017, and since then, several versions, including MobileNetV2 (Sandler et al., 2018) and V3 (Howard et al., 2019), have improved performance and efficiency.

The Inception CNN (Dongmei et al., 2020) is a CNN architecture developed by Google. It is designed to address the problem of choosing the optimal size of the convolutional filters, which are used to extract features from input images. This is achieved by using multiple filter sizes in parallel, ranging from small 1 × 1 filters to larger 5 × 5 filters, which capture information at different scales. These parallel filters are concatenated, allowing the network to capture fine and coarse details in the input image. The Inception architecture also includes a pooling operation that helps reduce the dimensionality of the feature maps, as well as auxiliary classifiers used during training to encourage the network to learn more robust features. The original Inception architecture was later refined and extended to create the Inception-V2 (Ioffe and Szegedy, 2015), Inception-V3 (Google, 2023), and Inception-V4 (Szegedy et al., 2016) architectures, which have achieved state-of-the-art results on several image recognition benchmarks.

Xception, as outlined in Fabien (2019), is a CNN architecture developed by Google. It was designed as an innovative approach inspired by the Inception architecture. In Inception, the original input was compressed using 1 × 1 convolutions, and distinct filters were applied to each depth space from each input space. Xception, on the other hand, reverses this process. It applies the filters to each depth map and subsequently compresses the input space using 1 × 1 convolutions across the depth. Additionally, the presence or absence of a nonlinearity following the initial operation varies. In the Inception architecture, both operations are accompanied by a ReLU nonlinearity. However, in Xception, no nonlinearity is introduced after the first operation.

VGG16 and VGG19 (Simonyan and Zisserman, 2014) are CNN architectures introduced by the Visual Geometry Group of Oxford University. In simple terms, VGG networks consist of convolutional layers with small 3 × 3 filters, then pooling layers, and finally, fully connected layers. The numbers 16 and 19 refer to the number of layers with weights in the network. The VGG architecture is widely used as a benchmark for image classification tasks because of its simplicity and robust performance. The network has been trained on the ImageNet dataset, which contains millions of labeled images across thousands of categories.

DenseNet121, referred to in Huang et al. (2017), is a deep neural network architecture that establishes dense connections between each layer and all other layers in a feedback manner. It consists of a basic convolution layer, a basic pooling layer, and multiple dense blocks with repeated convolutions. Transition layers interconnect these dense blocks. The network also includes a global average pooling layer for classification and an output layer. Overall, DenseNet-121 utilises 120 convolutions and four average pooling layers to facilitate its operations.

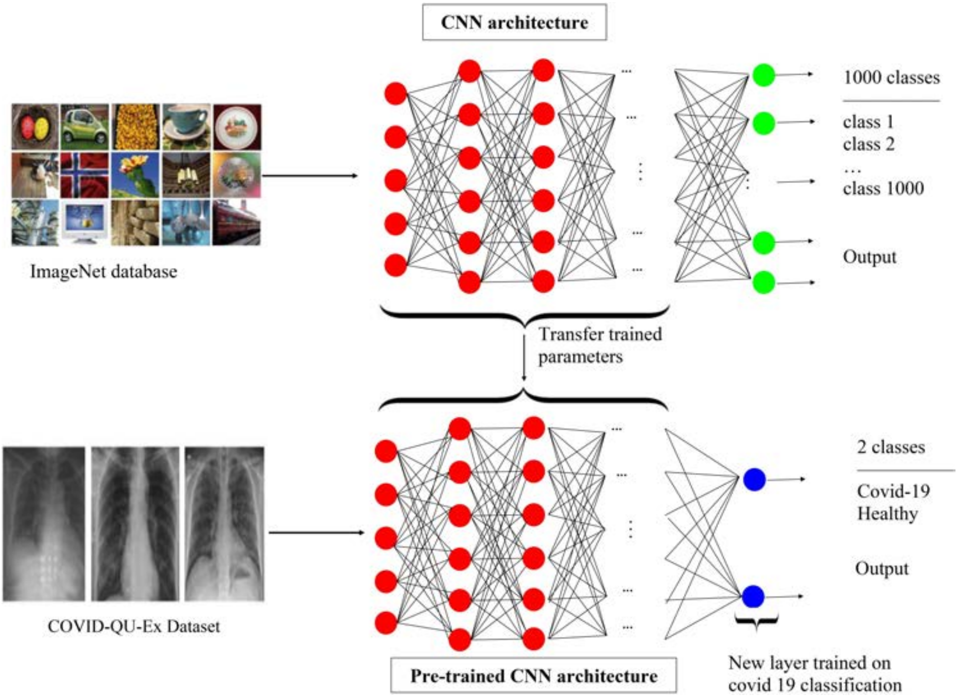
The NasNet (Zoph et al., 2018) is a highly efficient and accurate CNN-based architecture that uses neural architecture search to design neural network architectures automatically. It comprises four main building blocks, including normal cells, reduction cells, stem cells, and output cells. The network utilises skip connections and group convolution operations to improve its performance and efficiency on various computer vision tasks.

EfficientNet (Tan and Le, 2019) is a family of CNNs architectures developed by Google Brain in 2019. It was designed to achieve state-of-the-art accuracy while reducing the computational resources required. The EfficientNet B0 is the smallest version of this architecture, which balances accuracy and speed well.

The InceptionResNet (Szegedy et al., 2016) is a CNN architecture developed by Google. This architecture is a combination of Inception and ResNet50V2. The introduction of InceptionResNet aimed to address the challenge of vanishing gradients encountered in extremely deep neural networks. It consists of multiple Inception and ResNet50V2 modules stacked on each other. The Inception modules use a combination of 1 × 1, 3 × 3, and 5 × 5 convolutions to capture different scales of features. The ResNet50V2 modules consist of several convolutional layers with residual connections. The output of each module is then passed to the next module in the network. This architecture has achieved state-of-the-art results on benchmark datasets such as

ImageNet, CIFAR-10, and CIFAR-100. The architecture has also been used in various applications such as object detection, segmentation, and image captioning.

Figure 1 Representation of the transfer learning process in COVID-19's X-ray classification (see online version for colours)



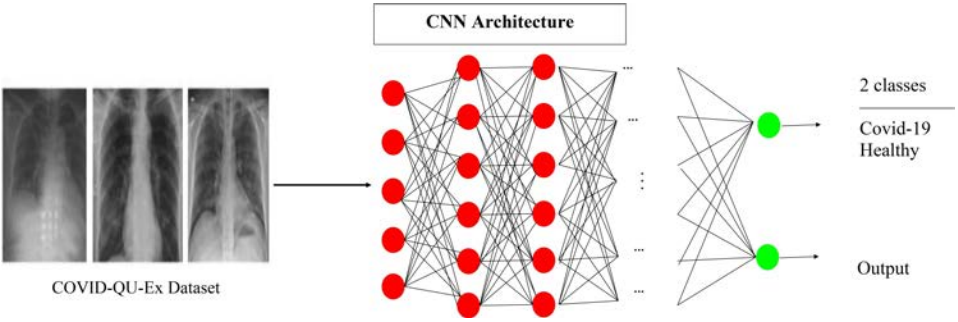
Notes: It represents the pre-trained CNN in its original state, capable of classifying 1,000 different classes from ImageNet. The last layer is replaced when using transfer learning. The dataset split 70% training, 20% validation, and 10% testing impacts model performance by ensuring sufficient training data while reserving validation data for hyperparameter tuning and a separate test set for unbiased evaluation. The pre-trained weights help accelerate convergence and improve generalisation, mainly when using a dataset of limited size.

The aforementioned deep CNN architectures have been employed for COVID-19 X-ray classification and have shown promising results, whether trained from scratch or using transfer learning. In the following section, we will empirically compare these deep CNN architectures under the same experimental framework using both approaches and the COVID-QU-Ex Dataset, one of the most extensive and up-to-date datasets available (Tahir et al., 2022).

3 Materials and methods

In this study, we perform a comprehensive analysis of 11 state-of-the-art deep CNN models: ResNet50V2, InceptionV3, MobileNetV2, VGG16, Xception, MobileNetV3, InceptionResnetV2, NasNetMobile, EfficientB0, and VGG19; for COVID-19's X-ray image classification with two configurations: transfer learning with fine-tuning and scratch learning. The above-listed deep CNN networks were taken from the library of TensorFlow (Abadi et al., 2016). In our study, weights are initialised randomly to train the model from scratch for full network training. The random initialisation of weight utilised only the architecture of the pre-trained model, not the weights trained on previous data. However, for transfer learning, the pre-trained network weights on the ImageNet (one of the most used image datasets in transfer learning) are kept based on the assumption that the pre-trained networks have already been trained well and fine-tuning is performed. In our study, we use the database COVID-QU-Ex, which consists of 33,920 CXR images, including 11,956 COVID-19, 11,263 non-COVID infections (viral or bacterial pneumonia), and 10,701 normal. The X-ray images are collected from different repositories and studies (Tahir et al., 2022), unlike other works that use some COVID-19 images from a single source to train the deep network from scratch or for fine-tuning. Figure 1 illustrates the process of using transfer learning. First, an architecture is pre-trained on the ImageNet dataset, as shown in the upper part of the figure. In the next step, we replaced the last layer (green neurons) responsible for classifying the classes (1,000 in ImageNet) with a layer that performs the classification of two classes (COVID-19 and normal) (blue neurons). The whole pre-trained CNN's layers were frozen, and only the dense layer was trained. Finally, we performed fine-tuning on the architecture using the COVID-QU-Ex dataset, as illustrated in the lower part of the image. Figure 2 shows the procedure of training the architecture from scratch, replacing the last layer with two neurons (green neurons) for classifying patients as COVID-19 or normal.

Figure 2 Classification process with a deep CNN architecture trained from scratch (see online version for colours)



Notes: Unlike transfer learning, where pre-trained weights are leveraged, training from scratch requires more data and computational resources. The dataset split – 70% training, 20% validation, and 10% testing plays a critical role in model performance. A large training set ensures the model learns meaningful representations, while the validation set helps prevent overfitting. However, training from scratch can be more sensitive to dataset size and variability, impacting generalisation performance.

The parameters of all architectures were set to default, and the input size for all architectures was 224 × 224. Images were rescaled, and no other preprocessing was applied. No data augmentation techniques were employed to focus our study exclusively on the deep architectures' performance. For each architecture, the last layer was replaced by a flattened layer and a dense layer with ReLU activation with the same number of neurons as the number of classes to classify (COVID-19 and normal). The stochastic gradient descent (SGD) optimiser was used with a learning rate of 0.001 and a momentum of 0.9. This configuration was chosen after preliminary experiments revealed that SGD with momentum led to stable convergence and reduced training fluctuations compared to adaptive optimisers. The loss function used was binary cross-entropy. A batch size of 32 was selected after testing various sizes (16, 32, 64, and 128), as it provided the best balance between computational efficiency and convergence stability. The dataset COVID-QU-Ex was divided into three subsets (see Table 1): 70% for training (15,860 images belonging to two classes), 20% for validation (4,532 images belonging to two classes), and 10% for testing (2,265 images belonging to two classes). No image augmentation techniques were used for training, and techniques like early stopping or any form of adaptive learning rate scheduling were employed to focus exclusively on the deep architecture's performance. To ensure efficient use of computational resources, each architecture was trained for 15 epochs. Limiting the number of epochs was based on the substantial time cost required to train each architecture. Although training deep architectures for a more extended period is expected, 15 epochs were enough for our purposes based on our experiments. We used accuracy, recall, precision, and F1-score to assess the performance of the evaluated deep CNN architectures because these metrics are those used in the state of the art for the classification of COVID-19. In the evaluation process, the number of instances correctly predicted as positive is denoted as true positive (TP). At the same time, false negative (FN) represents the number of cases incorrectly predicted as negative. True negative (TN) refers to the number of negative cases correctly predicted, and false positive (FP) represents the number of negative cases incorrectly predicted as positive. Equations (1), (2), (3) and (4) were utilised to calculate the evaluation metrics based on the values of TP, TN, FP, and FN.

Table 1 Distribution of COVID-QU-Ex dataset for training, validation, and testing

Dataset split	*Percentage*	*Number of images*
Training	70%	15,860
Validation	20%	4,532
Testing	10%	2,265

Sure of correctly classified COVID-19 cases and is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

Precision is defined as the measure of correctly classified instances in truly positive instances and is defined as:

$$Precision = \frac{TP}{TP + FN} \tag{2}$$

Accuracy is the measure of correctly classified instances divided by the total number of test cases and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

F1-score or F1 is defined as the weighted average of precision and recall and is defined as:

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{4}$$

We ran the experiments using Kaggle's Nvidia P100, which has 16 GB of VRAM, 13GB of RAM, and a 2-core Intel Xeon CPU. These computational resources limited the number of epochs and image size used during training, as some models are very resource-intensive due to the large number of trainable parameters. Therefore, only the previously mentioned configuration, which all models could use, was selected to ensure a fair comparison and avoid a deep architecture outperforming others due to the leverage of resources.

4 Results

In this section, we present the results of evaluating the deep CNN architectures presented in the state-of-the-art for their ability to classify healthy and sick individuals in COVID-19 datasets. We compare these architectures under different metrics when trained using transfer learning with pre-trained architectures versus training from scratch. We assessed DenseNet121, InceptionV3, MobileNetV1, MobileNetV2, MobileNetV3, VGG16, VGG19, Xception, InceptionResNetV2, NasNetMobile, ResNet50V2, and EfficientNetB0. Table 2 compares the number of parameters in each evaluated architecture under two approaches: training from scratch and transfer learning. The second column specifies the number of trainable parameters when the architecture is trained from scratch, the third column indicates the number of trainable parameters when applying transfer learning. Table 3 presents the classification results; each cell shows the classification quality reached by each architecture in the test set of the dataset, COVID-QU-Ex (as commented above). Table 3 is divided into two parts; the first four columns show the classification results, with the metrics accuracy, F1-score, recall, and precision when the deep architecture is trained from scratch; the last four columns show the classification results for the same four metrics when the deep architecture is trained using transfer learning with fine-tuning.

4.1 Discussion

Based on the results obtained from Table 3, it is evident that training deep CNN architectures from scratch consistently outperformed the transfer learning approach using ImageNet weights in all evaluated metrics. This indicates that training the deep architectures on the COVID-QU-Ex dataset from scratch is more effective in accurately classifying COVID-19 patients. We can observe that employing transfer learning with fine-tuning, commonly used when a few images are available, was a good alternative. However, more data are currently available (with greater diversity) in the COVID-QU-Ex dataset; thus, as expected, using deep architectures and training them from scratch got better results.

Table 2 Comparison of the number of parameters of the tested architectures using transfer learning vs. trained from scratch

Architecture	*Trainable parameters (scratch)*	*Trainable parameters (TL)*
DenseNet121	6,955,906	2,050
InceptionV3	21,915,810	147,458
MobileNetV2	2,387,714	163,842
VGG16	16,812,032	2,097,346
Xception	21,007,658	200,706
MobileNetV3	1,517,850	2,050
InceptionResNetV2	54,352,994	76,802
NasNetMobile	53,643,522	4,269,716
ResNet50V2	23,519,360	200,706
EfficientNet B0	4,132,990	125,442
VGG19	20,074,560	50,178

Table 3 Comparison of the deep CNN architectures using transfer learning vs. training from scratch

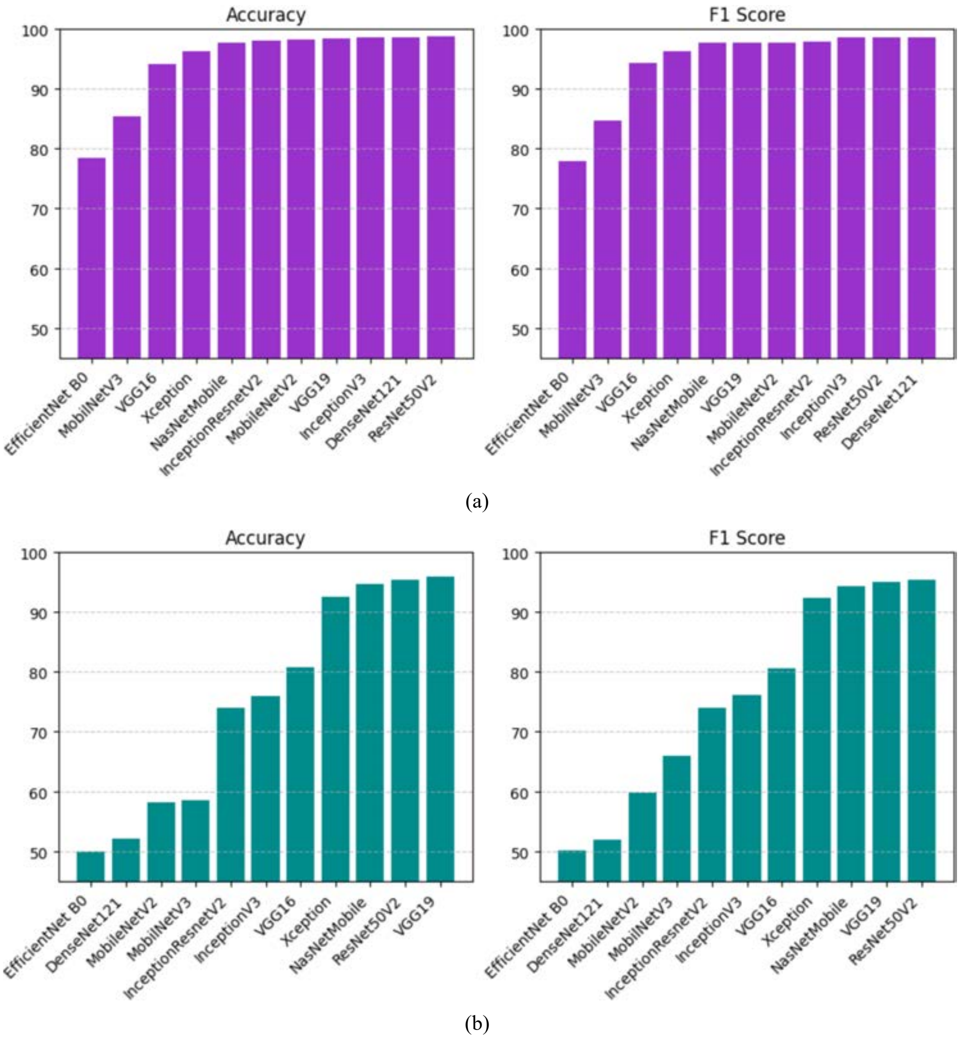
Architecture	*From scratch*				*Transfer learning*			
	Accuracy	*F1-score*	*Recall*	*Precision*	*Accuracy*	*F1-score*	*Recall*	*Precision*
DenseNet121	98.69	*98.70*	98.63	*98.79*	52.23	52.07	50.76	53.51
InceptionV3	98.58	98.59	98.60	98.57	76.05	76.08	76.07	71.06
MobileNetV2	98.25	97.76	98.86	98.72	58.23	59.85	60.90	58.90
VGG16	94.25	94.33	93.81	94.88	80.81	80.64	80.19	81.16
Xception	96.31	96.34	96.29	96.25	92.50	92.32	92.32	92.32
MobilNetV3	85.35	84.66	88.24	81.44	58.65	65.92	86.06	53.47
InceptionResnetV2	98.10	98.00	97.70	98.20	74.01	74.05	72.08	74.09
NasNetMobile	97.81	97.69	97.71	97.58	94.81	94.42	92.59	95.38
ResNet50V2	*98.81*	98.68	98.72	98.58	95.51	*95.52*	*94.91*	95.59
EfficientNet B0	78.40	77.86	76.74	79.97	50.00	50.20	50.20	50.20
VGG19	98.47	97.69	*99.46*	96.02	*95.90*	95.12	92.46	*98.05*

Notes: The best results for each metric appear in italic.

To better appreciate which CNNs obtain the best results in both accuracy and F1-score. In Figure 3 we graph on the y-axis the accuracy (F1-score) of each of the CNNs that appear on the x-axis. In this graph, we have ordered the CNNs from lowest to highest accuracy (F1-score) in such a way that the CNNs that appear furthest to the right are the ones that obtain the highest values. Figure 3(a) shows the results of the CNNs trained from scratch, while Figure 3(b) shows the results of the CNNs trained using transfer learning and fine-tuning.

From Figure 3(a), we can see that when trained from scratch, ResNet50V2, DenseNet121, and InceptionV3 are the CNNs that appear most right in both accuracy and F1-score graphs, i.e., those that got the best classification results. Specifically, ResNet50V2 stands out with an accuracy of 98.81% and an F1-score of 98.68%. It is closely followed by DenseNet121, which achieves 98.69% accuracy and a 98.70% F1-score, and InceptionV3, which achieves an accuracy of 98.58% and an F1-score of 98.59%. On the other hand, Figure 3(b) shows that when employing transfer learning, the CNNs VGG19 and ResNet50V2 appear most right in both accuracy and F1-score graphs, i.e., they got the best classification results. Specifically, VGG19 exhibits superior performance, with an accuracy of 95.90% and an F1-score of 95.12%, alongside ResNet50V2, with an accuracy of 95.51% and an F1-score of 95.52%.

Figure 3 Accuracy and F1-score values reached by the deep CNN architectures, (a) accuracy and F1-score values reached by the CNNs by training from scratch (b) accuracy and F1-score values reached by the CNNs by training with transfer learning and fine-tuning (see online version for colours)



(a)

(b)

The results shown in Table 3 that deep architectures such as EfficientNet B0, DenseNet121, MobilNetV3, InceptionResnetV2, and InceptionResnetV3 got poor results for COVID-19's X-ray image classification when the configuration transfer learning with fine-tuning was employed. Nevertheless, deep architectures such as VGG19 and ResNet50V2 are good alternatives for this configuration, which can be very useful when there are insufficient computational resources to train these dense architectures from scratch. Something that does not happen with the NasNetMobile architecture, which, although it gets good classification results, has a large number of parameters (twenty times more parameters than ResNet50V2 and Xception and eighty-five times more parameters than VGG19).

Several limitations must be considered. EfficientNet B0 underperformed, likely due to the small size of its architecture, making it unsuitable for capturing the intricate patterns in COVID-19 X-ray images. MobileNetV3 and InceptionResNetV2 showed moderate results, potentially due to their limited depth, affecting their ability to detect subtle abnormalities. Despite demonstrating good performance when trained from scratch, DenseNet121 and NasNet demand significant computational resources, limiting their practical applications in resource-constrained settings. Additionally, although VGG19 performed well in the transfer learning configuration, VGG16 and VGG19 require high memory and long training times, hindering rapid deployment.

From our experiments, we can conclude that for COVID-19's X-ray image classification, among all the architectures tested, the best one is VGG19 since it obtained the highest percentages in accuracy (which gives us the total percentage of correctly classified patients) and the highest percentages in precision (which is the percentage of patients correctly diagnosed with COVID-19). These results are comparable with those obtained with the configuration training from scratch because the different metrics do not vary more than 5% below training from scratch. As seen in Table 2, among all the other deep CNN architectures that got good results, the VGG19 architecture needs to train the fewest number of parameters with transfer learning.

Our experiments eliminate the problem mentioned by the authors of different papers in the related work: little variability in the used datasets; for instance, Kassania et al. (2021) and Khan et al. (2022) highlighted that despite the good classification results reached on COVID-19's X-ray image classification their study should be considered early studies due to a lack of data variety; the dataset used in their experiments consisted of patients from the same hospital with similar demographic characteristics, which hindered the generalisability of transfer learning-based training of deep learning networks. Consequently, overfitting was observed in most of the architectures explored in the literature. In this regard, in our study, we employed a dataset (the COVID-QU-Ex dataset) that combined multiple public datasets, thus increasing the variability of the data. Under these variability conditions in the data, from our results, we can see that ResNet50V2 is the deep architecture that appears among the best ones when trained from scratch and through transfer learning with fine-tuning. Therefore, our study allows us to conclude that ResNet50V2 can effectively be employed for COVID-19's X-ray image classification in any of these configurations. The performance of ResNet50V2 when trained from scratch highlights the potential of customised training for COVID-19 detection, especially in well-resourced healthcare settings. However, the high performance of VGG19 using transfer learning suggests that hospitals with limited computational resources can still achieve accurate diagnostics by leveraging pre-trained networks. These findings could aid radiologists and clinicians in selecting appropriate

AI-based diagnostic tools tailored to their available resources. Furthermore, in current medical practice, deploying these models could streamline diagnostic workflows, reduce radiologist workload, and provide faster patient triaging, essential during peak infection periods.

5 Conclusions

In the medical field, it is crucial to consider methods based on deep learning when dealing with classification problems. Due to the black-box nature of deep learning methods, it is necessary to ensure good quality results of the classifier. Typically, a large dataset is used for this purpose. Transfer learning has yielded favourable results for various deep CNN architectures. However, as our study reveals, not all architectures perform well when utilising this technique for COVID-19's X-ray image classification. Hence, it is crucial to choose the appropriate deep CNN architecture for a specific task and carefully evaluate its performance, particularly in medical applications where the correct classification is vital. Our study of 11 state-of-the-art deep CNN models for COVID-19's X-ray image classification with the two configurations more studied in the literature, transfer learning with fine-tuning and training from scratch, concludes that the ResNet50V2 architecture is the best option for COVID-19's X-ray image classification, in terms of accuracy and F1-score, when trained from scratch and through transfer learning with fine-tuning. However, if computational resources are limited, one can opt for VGG19 using transfer learning through the use of ImageNet. These findings are valuable insights for researchers and practitioners working on COVID-19 classification using deep-learning CNNs, opening new opportunities to expand access to fast and accurate diagnostics.

Since our study focused on training from scratch or transfer learning on different deep CNN architectures and did not include early stopping and adaptive learning rate scheduling techniques, the impact of these techniques on the deep CNN architectures for COVID-19's X-ray image classification will be performed as future work. It is also important to evaluate how adjusting the number of layers to train in transfer learning affects model performance, especially considering the availability of images for training. Adjusting this parameter can significantly improve results, allowing the model to better adapt to different data volumes. Additionally, generative adversarial networks (GANs) could be used to create synthetic images, which would increase the amount of available data and allow for comparison of current models with others, identifying which one performs best with a larger number of images. The above is especially relevant, as some models, due to the number of trainable parameters, may find that the dataset used was not large enough. Future research could further enhance model interpretability by incorporating explainability techniques such as gradient-weighted class activation mapping (Grad-CAM). Applying Grad-CAM visualisations to models like ResNet50V2 and VGG19 would allow a deeper understanding of how these architectures make classification decisions by highlighting the most relevant radiological features influencing predictions. This approach could help identify whether the models rely on clinically meaningful regions of the X-ray images or are affected by dataset biases. Additionally, integrating explainability frameworks could improve trust and adoption of deep learning models in real-world medical settings, enabling clinicians to validate AI-driven diagnostic recommendations more effectively. Future work may also explore hybrid approaches, combining CNNs with transformer-based models to enhance feature extraction while maintaining interpretability through visualisation techniques. Furthermore, an important area for future research involves analysing misclassified images to better understand where models fail. Examining whether errors occur more frequently in mild COVID-19 cases or if models tend to confuse COVID-19 with other pneumonia types could provide critical insights into model limitations. While this analysis was considered during the study, it was ultimately not included in the final version to maintain focus on the broader performance comparison across architectures.

Acknowledgements

This research was partially supported by the National Council of Humanities, Sciences, Technologies, and Innovation of Mexico (CONAHCyT) through its graduate study scholarship program.

Declarations

The authors have no competing interests to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject or materials discussed in this manuscript.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015) *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* [online] http://tensorflow.org (accessed March 2024).

Abdullah, M., Abrha, F.B., Kedir, B. and Tagesse, T.T. (2024) 'A hybrid deep learning CNN model for COVID-19 detection from chest X-rays', *Heliyon*, Vol. 10, No. 5, p.e26938.

Alsattar, H.A., Qahtan, S., Zaidan, A.A., Deveci, M., Martinez, L., Pamucar, D. and Pedrycz, W. (2024) 'Developing deep transfer and machine learning models of chest X-ray for diagnosing covid-19 cases using probabilistic single-valued neutrosophic hesitant fuzzy', *Expert Systems with Applications*, Vol. 236, p.121300.

Aslani, S. and Jacob, J. (2023) 'Utilisation of deep learning for COVID-19 diagnosis', *Clinical Radiology, Special Issue Section: Artificial Intelligence and Machine Learning*, Vol. 78, No. 2, pp.150–157.

Bhosale, Y.H. and Patnaik, K.S. (2023) 'Application of deep learning techniques in diagnosis of COVID-19 (coronavirus): a systematic review', *Neural Processing Letters*, Vol. 55, No. 3, pp.3551–3603.

Cao, D.M., Amin, M.S., Islam, M.T., Ahmad, S., Haque, M.S., Sayed, M.A., Rahman, M.M., Koli, T., Ayon, E.H. and Nobe, N. (2023) 'Deep learning-based COVID-19 detection from chest X-ray images: a comparative study', *Journal of Computer Science and Technology Studies*, Vol. 5, No. 4, pp.132–141.

Constantinou, M., Exarchos, T., Vrahatis, A.G. and Vlamos, P. (2023) 'COVID-19 classification on chest X-ray images using deep learning methods', *International Journal of Environmental Research and Public Health*, Vol. 20, No. 3.

Deng, J., Dong, W., Socher, R., Li, L-J., Li, K. and Fei-Fei, L. (2009) 'ImageNet: a large-scale hierarchical image database', in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp.248–255.

Dongmei, Z., Ke, W., Hongbo, G., Peng, W., Chao, W. and Shaofeng, P. (2020) 'Classification and identification of citrus pests based on InceptionV3 convolutional neural network and migration learning', *2020 International Conference on Internet of Things and Intelligent Applications (ITIA)*, pp.1–7.

Eskandarian, R., Alizadehsani, R., Behjati, M., Zahmatkesh, M., Sani, Z., Haddadi, A., Kakhi, K., Roshanzamir, M., Shoeibi, A., Hussain, S., Khozeimeh, F., Darbandy, M., Joloudari, J.H., Lashgari, R., Khosravi, A., Nahavandi, S. and Islam, S.M.S. (2023) 'Identification of clinical features associated with mortality in COVID-19 patients', *Operations Research Forum*, Vol. 4, No. 3, p.2035.

Fabien, M. (2019) *XCeption Model and Depthwise Separable Convolutions*, March [online] https://maelfabien.github.io/deeplearning/xception/# (accessed March 2024).

Farooq, M.S. and Hafeez, A. (2020) *COVID-ResNet: A Deep Learning Framework for Screening of COVID-19 from Radiographs*, arXiv preprint arXiv::2003.14395.

Fernández-Miranda, P.M., Fraguela, E.M., Álvarez de Linera-Alperi, M., Cobo, M., del Barrio, A.P., González, D.R., Vega, J.A. and Iglesias, L.L. (2024) 'A retrospective study of deep learning generalization across two centers and multiple models of X-ray devices using COVID-19 chest X-rays', *Scientific Reports*, June, Vol. 14, p.14657.

Google (2023) *Inception V3 Advanced Guide* [online] https://cloud.google.com/tpu/docs/inception-v3-advanced?hl=es-419 (accessed March 2023).

Hassan, E., Shams, M.Y., Hikal, N.A. and Elmougy, S. (2024) 'Detecting COVID-19 in chest CT images based on several pre-trained models', *Multimedia Tools and Applications*, Vol. 83, No. 24, pp.65267–65287.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770–778.

Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017) *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*, April, arXiv preprint arXiv:1704.04861.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H. and Le, Q. (2019) 'Searching for MobileNetV3', *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.1314–1324.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) 'Densely connected convolutional networks', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Ioffe, S. and Szegedy, C. (2015) 'Batch normalization: accelerating deep network training by reducing internal covariate shift', *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37, pp.448–456.

Kassania, S.H., Kassanib, P.H., Wesolowskic, M.J., Schneidera, K.A. and Detersa, R. (2021) 'Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning based approach', *Biocybernetics and Biomedical Engineering*, Vol. 41, No. 3, pp.867–879.

Khan, S.H., Sohail, A. and Khan, A. (2022) 'COVID-19 detection in chest X-ray images using a new channel boosted CNN', *Diagnostics*, Vol. 12, No. 1, p.267.

Khattab, R., Abdelmaksoud, I.R. and Abdelrazek, S. (2024) 'Automated detection of COVID-19 and pneumonia diseases using data mining and transfer learning algorithms with focal loss from chest X-ray images', *Applied Soft Computing*, Vol. 162, p.111806.

Mooney, P. (2018) *Chest X-Ray Images (Pneumonia)*, March [online] https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia (accessed March 2023).

Paul, S.G., Saha, A., Biswas, A.A., Zulfiker, M.S., Arefin, M.S., Rahman, M.M. and Reza, A.W. (2023) 'Combating COVID-19 using machine learning and deep learning: applications, challenges, and future perspectives', *Array*, Vol. 17, p.100271.

Ramadhan, A.A. and Baykara, M. (2022) 'A novel approach to detect COVID-19: enhanced deep learning models with convolutional neural networks', *Applied Sciences*, Vol. 12, No. 9, p.9325.

Reddy, A.S.K., Rao, K.N.B., Soora, N.R., Shailaja, K., Kumar, N.C.S., Sridharan, A. and Uthayakumar, J. (2023) 'Multi-modal fusion of deep transfer learning based covid-19 diagnosis and classification using chest X-ray images', *Multimedia Tools and Applications*, Vol. 82, No. 8, pp.12653–12677.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L-C. (2018) 'MobilenetV2: inverted residuals and linear bottlenecks', *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June, pp.4510–4520.

Simonyan, K. and Zisserman, A. (2014) *Very Deep Convolutional Networks for Large-Scale Image Recognition*, September, arXiv 1409.1556.

Singh, T., Mishra, S., Kalra, R., Satakshi, Kumar, M. and Kim, T. (2024) 'COVID-19 severity detection using chest X-ray segmentation and deep learning', *Scientific Reports*, August, Vol. 14, p.19846.

Stein, A., MD, Wu, C. and Carr, C. (2018) *RSNA Pneumonia Detection Challenge* [online] https: //kaggle.com/competitions/rsna-pneumonia-detection-challenge (accessed March 2023).

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A. (2016) 'Inception-V4, Inception-ResNet and the impact of residual connections on learning', *Proceedings of the ... AAAI Conference on Artificial Intelligence*, Vol. 31, No. 2.

Tahir, A.M., Chowdhury, M.E.H., Qiblawey, Y., Khandakar, A., Rahman, T., Kiranyaz, S., Khurshid, U., Ibtehaz, N., Mahmud, S. and Ezeddin, M. (2022) *Covid-Qu-Ex Dataset*, Kaggle [online] https://doi.org/10.34740/KAGGLE/DSV/3122958.

Tan, M. and Le, Q.V. (2019) *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, ArXiv, abs/1905.11946.

Ukwuoma, C.C., Cai, D., Heyat, M.B.B., Bamisile, O., Adun, H., Al-Huda, Z. and Al-Antari, M.A. (2023) 'Deep learning framework for rapid and accurate respiratory COVID-19 prediction using chest X-ray images', *Journal of King Saud University – Computer and Information Sciences*, Vol. 35, No. 7, p.101596.

Wang, Z., Zhang, K. and Wang, B. (2022) 'Detection of COVID-19 cases based on deep learning with X-ray images', *Electronics*, Vol. 11, No. 10, p.3511.

Wu, Y-H., Yeh, I-J., Phan, N.T.S., Yen, M-C., Hung, J-H., Chiao, C-C., Chen, C-F., Sun, Z., Hsu, H.P., Wang, C-Y. and Lai, M.C. (2021) 'Gene signatures and potential therapeutic targets of Middle East respiratory syndrome coronavirus (MERS-CoV)-infected human lung adenocarcinoma epithelial cells', *Journal of Microbiology Immunology and Infection*, Vol. 54, No. 3, pp.845–857.

Zoph, B., Vasudevan, V., Shlens, J. and Le, Q.V. (2018) 'Learning transferable architectures for scalable image recognition', *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June.